Extreme-scale AI computing with Cerebras

A Hock *
Cerebras Systems

* andy@cerebras.net

Argonne Training Program on Extreme-Scale Computing (ATPESC)
27 July 2020

Agenda

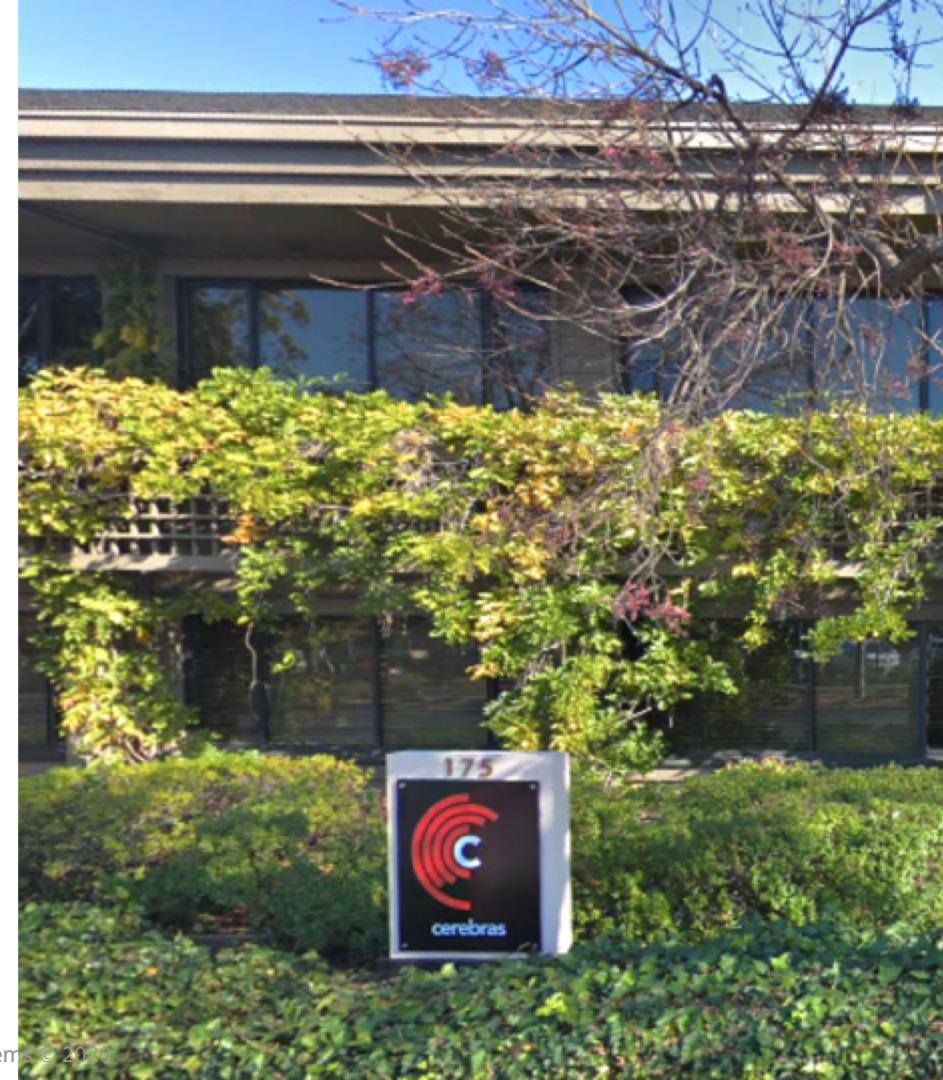
Introduction
Programming the CS-1
Value for AI research
Conclusions

Cerebras Systems © 2020

Cerebras Systems

Transform the compute landscape
Radically accelerate AI

Founded 2016
200+ world-class engineers
Hardware, software, ML/AI research



cerebras

AI has **massive potential**,
but is **compute-limited today**.

We **need a new compute solution**
to accelerate deep learning.

Cerebras Systems © 2020

Compute for deep learning is hard

Massive compute
- Billions-trillions of ops per sample
- Millions-billions of samples per training
- Peta-exa scale compute

Memory footprint
- GB+ weights, TB+ datasets

High bandwidth communication
- Many neurons, many topologies

Often leads to days-weeks training time

Cerebras Systems © 2020

Training takes too long

Experiment Turnaround Time and Research Productivity

- **Minutes, Hours**:
  - **Interactive research! Instant gratification!**
- **1-4 days**
  - Tolerable
  - Interactivity replaced by running many experiments in parallel
- **1-4 weeks**
  - High value experiments only
  - Progress stalls
- **>1 month**
  - Don't even try

Ref: Dean et al. 2017. Google Brain.

Cerebras Systems © 2020

Existing processors can’t keep up

Traditional general purpose processors were not built for this work

- Compute cores, silicon not optimized for deep learning
- Not built for high bandwidth communication, memory access

Scaling out over many such processors poses significant challenges

- Inefficient scaling, silicon utilization
- Scale-out implementations brittle, difficult to program

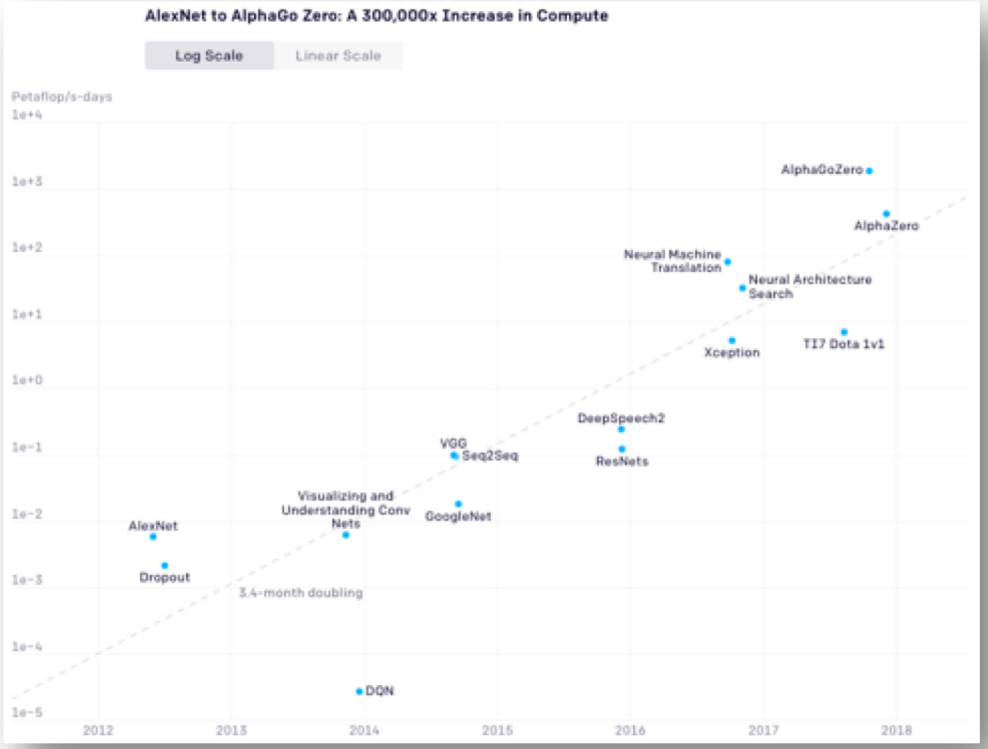
cerebras

Cerebras Systems © 2020

...and the challenge is growing.

Compute for largest AI training tasks has increased by 300,000x since 2012.

A 3.4-month exponential doubling time. By comparison, Moore's law had a 2-year doubling time.



Ref: Amodei & Hernandez 2018. OpenAI.

Cerebras Systems © 2020

The right solution for AI compute

Many cores **optimized for sparse linear algebra**

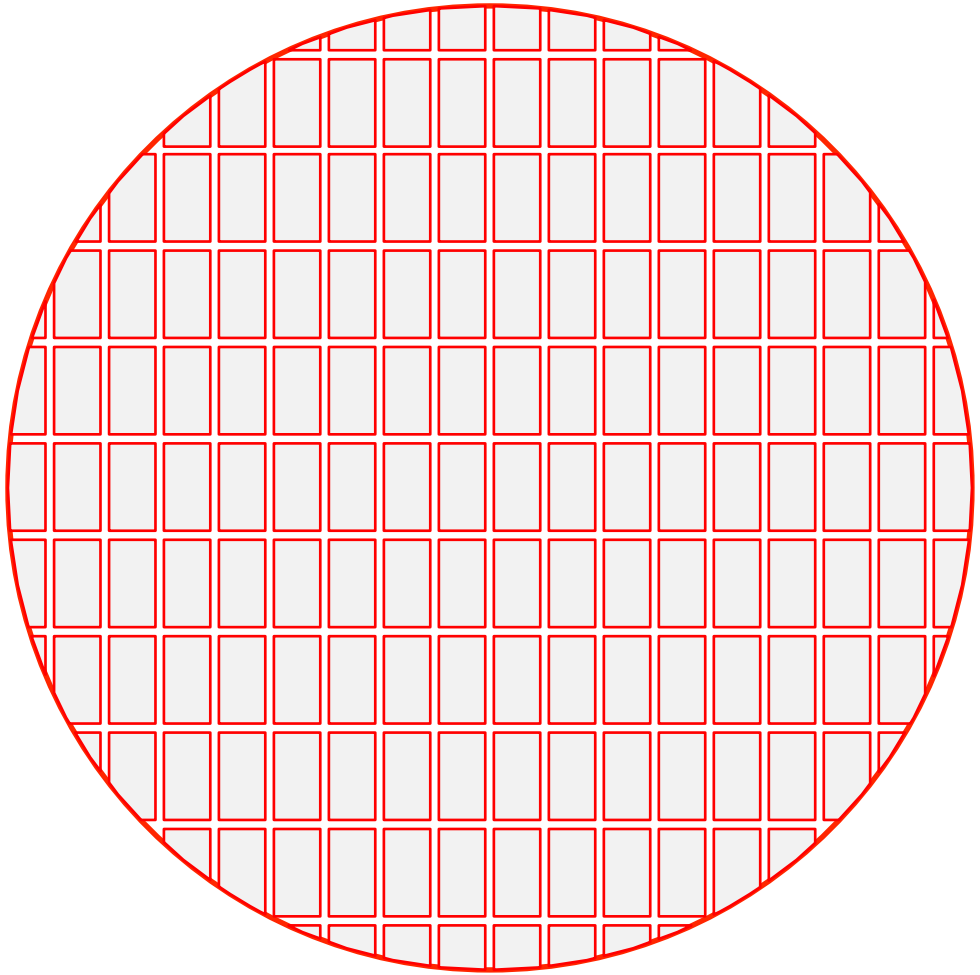
Memory tightly coupled to compute

High bandwidth **communication**

Programmable with today's **ML frameworks**

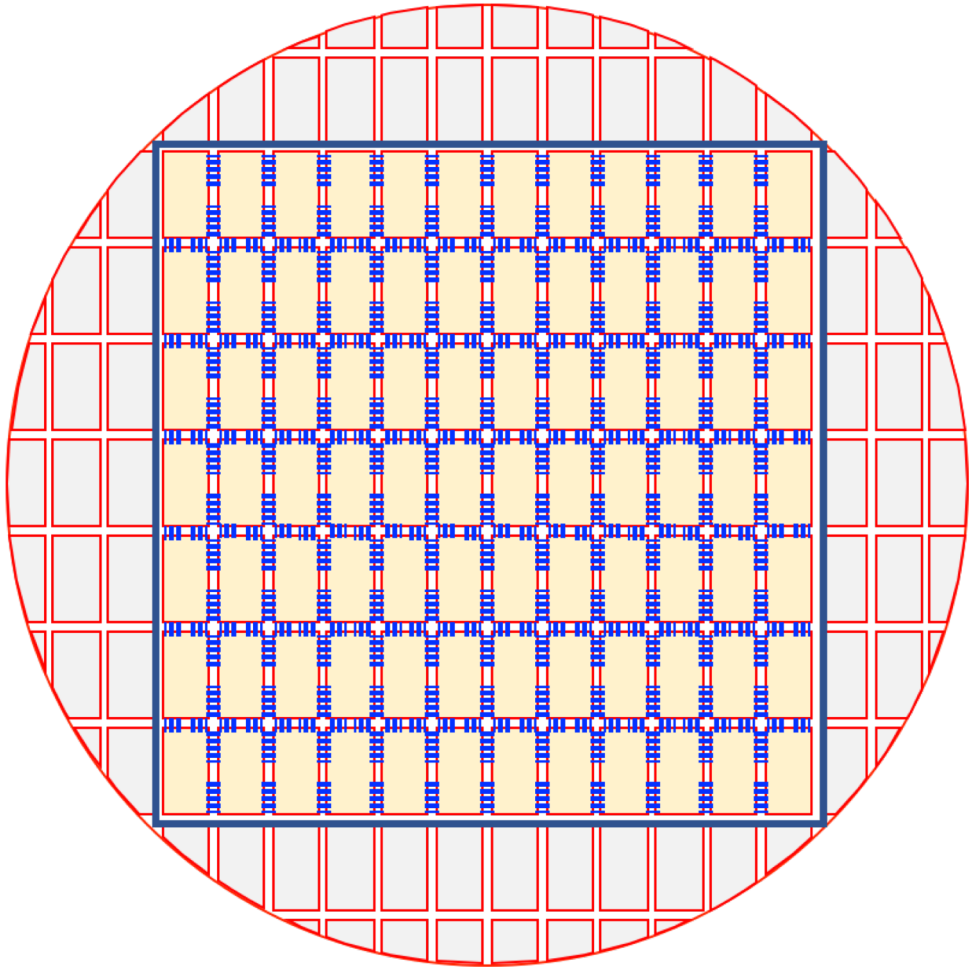
cerebras

Cerebras Systems © 2020



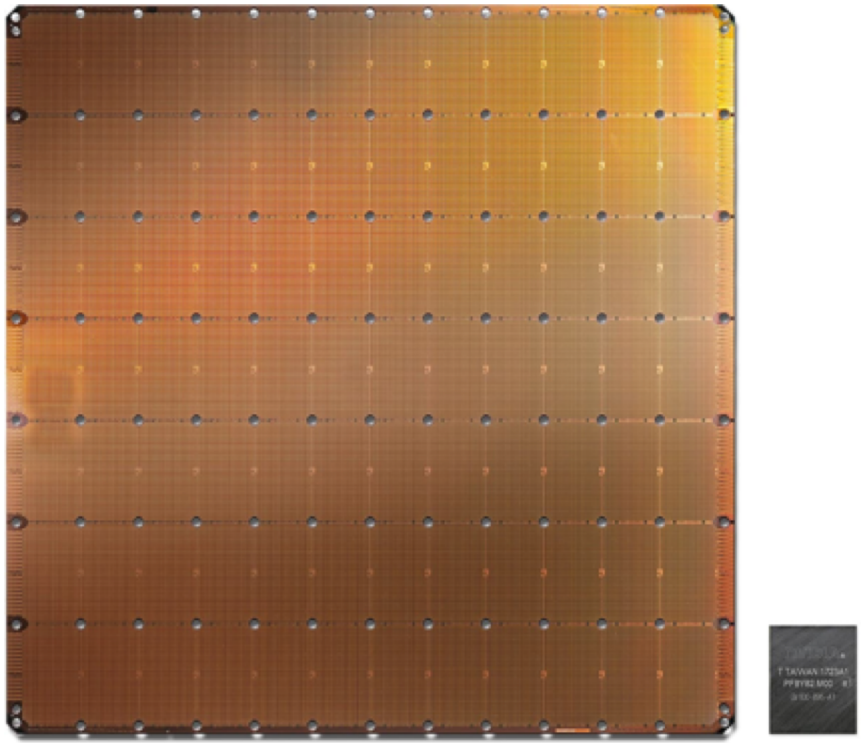
Cerebras Systems © 2020

cerebras



Cerebras Systems © 2020

The Cerebras Wafer-Scale Engine



The world's largest chip and most powerful AI engine.

Designed from the ground-up to deliver orders of magnitude performance gain for deep learning.

- 215 x 215 mm, 1.2 trillion transistor chip
- 400,000 cores
- 18 GB on-chip SRAM
- 100 Pb/s interconnect

cerebras

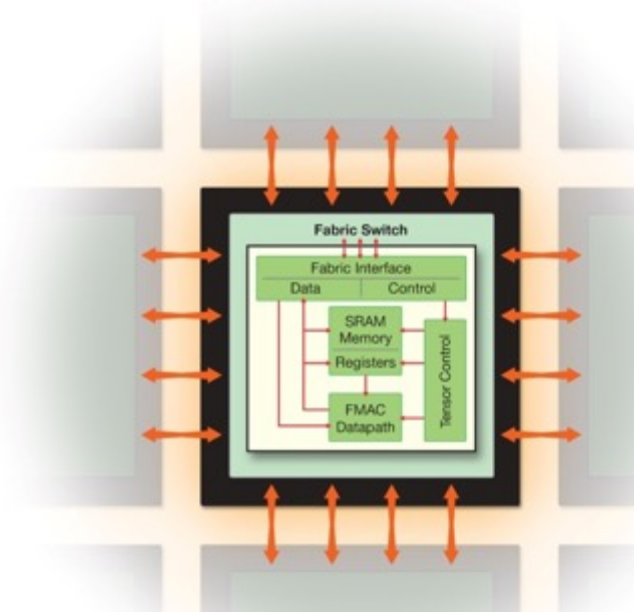
Cerebras Systems © 2020

	Cerebras WSE	Largest GPU	Cerebras Advantage
Chip size	46,225 mm^2	815 mm^2	56.7 X
Cores	400,000	5,120	78 X
On chip memory	18 Gigabytes	6 Megabytes	3,000 X
Memory bandwidth	9 Petabytes/S	900 Gigabytes/S	10,000 X
Fabric bandwidth	100 Petabits/S	300 Gigabits/S	33,000 X

The CS WSE architecture is built for deep learning

AI-optimized **compute**

- Fully-programmable core, ML-optimized extensions
- Dataflow architecture optimized for sparse, dynamic workloads

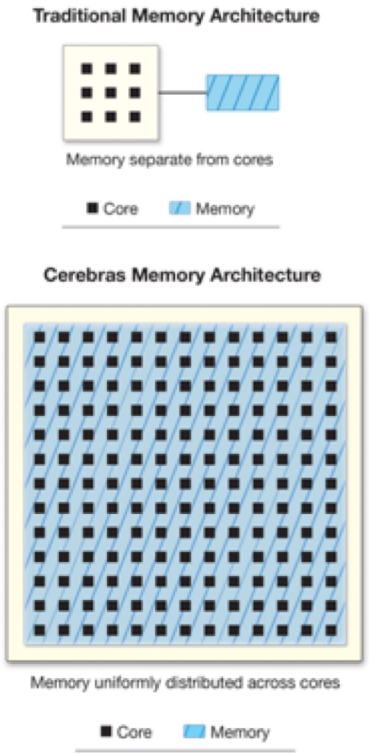


Cerebras Systems © 2020

The CS WSE architecture is built for deep learning

AI-optimized **compute**
- Fully-programmable core, ML-optimized extensions
- Dataflow architecture optimized for sparse, dynamic workloads

AI-optimized **memory**
- Traditional memory architectures shared memory far from compute
- The right answer is distributed, high performance, on-chip memory



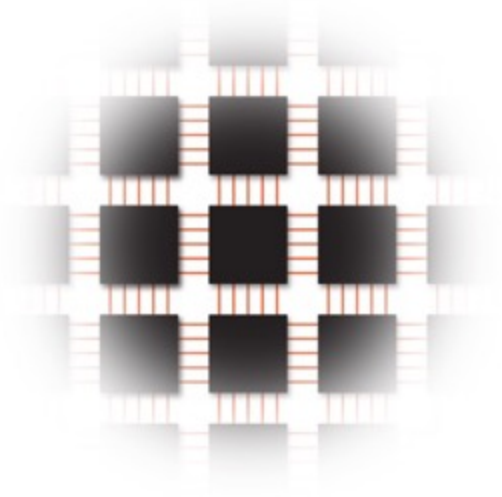
Cerebras Systems © 2020

The CS WSE architecture is built for deep learning

AI-optimized **compute**
- Fully-programmable core, ML-optimized extensions
- Dataflow architecture optimized for sparse, dynamic workloads

AI-optimized **memory**
- Traditional memory architectures shared memory far from compute
- The right answer is distributed, high performance, on-chip memory

AI-optimized **communication**
- High bandwidth, low latency cluster-scale networking on chip
- Fully-configurable to user-specified topology



cerebras

Cerebras Systems © 2020

The CS WSE architecture is built for deep learning

AI-optimized **compute**
- Fully-programmable core, ML-optimized extensions
- Dataflow architecture optimized for sparse, dynamic workloads

AI-optimized **memory**
- Traditional memory architectures shared memory far from compute
- The right answer is distributed, high performance, on-chip memory

AI-optimized **communication**
- High bandwidth, low latency cluster-scale networking on chip
- Fully-configurable to user-specified topology

Together, orders of magnitude performance and efficiency gain
Native model parallel execution
Full utilization at small batch, accelerated sparse compute

cerebras

Cerebras Systems © 2020

Systems-First Approach

The **challenge**: how to harness this performance to achieve orders of magnitude gain?

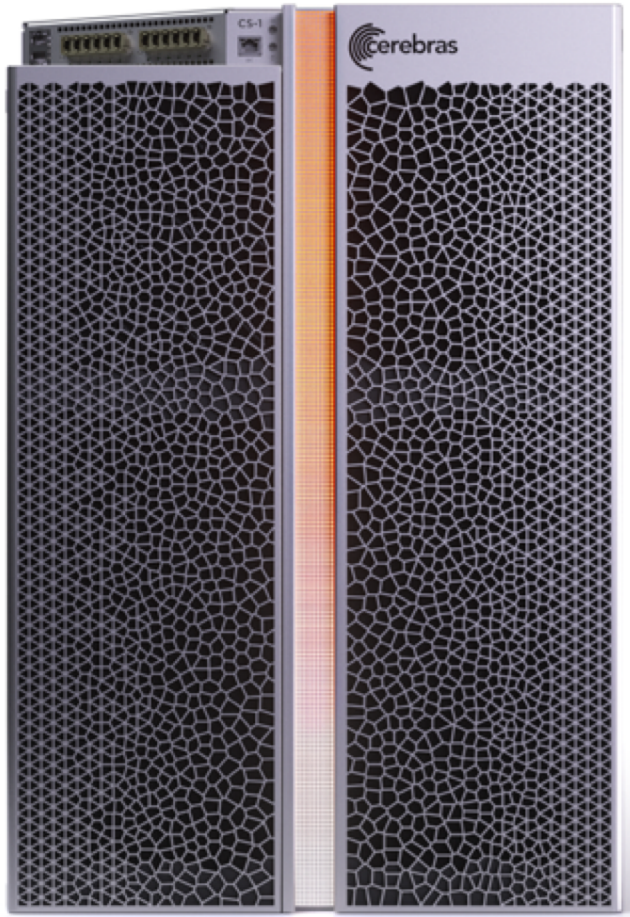
e.g. package, power, cool, make it easy to deploy?

Impossible with off-the-shelf / legacy technologies.

(You wouldn't put a racecar engine in a minivan and expect to win a Grand Prix.)

Orders of magnitude performance gain **requires systems-level thinking**.

Cerebras Systems © 2020



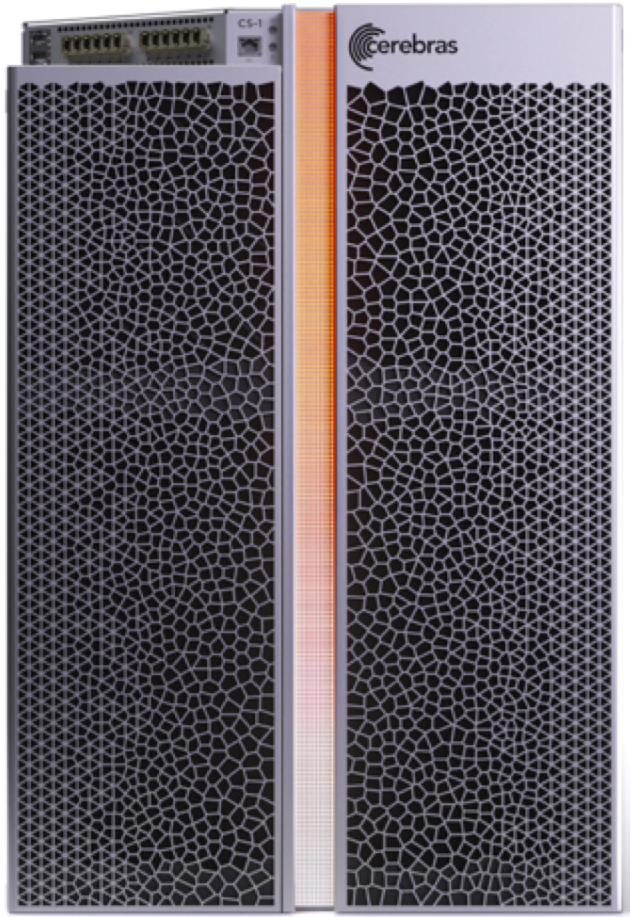
The Cerebras CS-1

The world's most powerful AI computer

A **full solution** in a single system:
- Powered by the WSE
- Programmable via TF, other frameworks
- Install, deploy easily into a standard rack

cerebras

ystems © 2020



The Cerebras CS-1

The world's most powerful AI computer

A **full solution** in a single system:
- Powered by the WSE
- Programmable via TF, other frameworks
- Install, deploy easily into a standard rack

15 RU standard rack-compliant server
1.2 Tbps I/O via 12x100GbE
20 kW power, air-cooled

Replace racks of legacy general purpose servers with a single system <1 rack

cerebras

ystems © 2020

The Cerebras CS-1

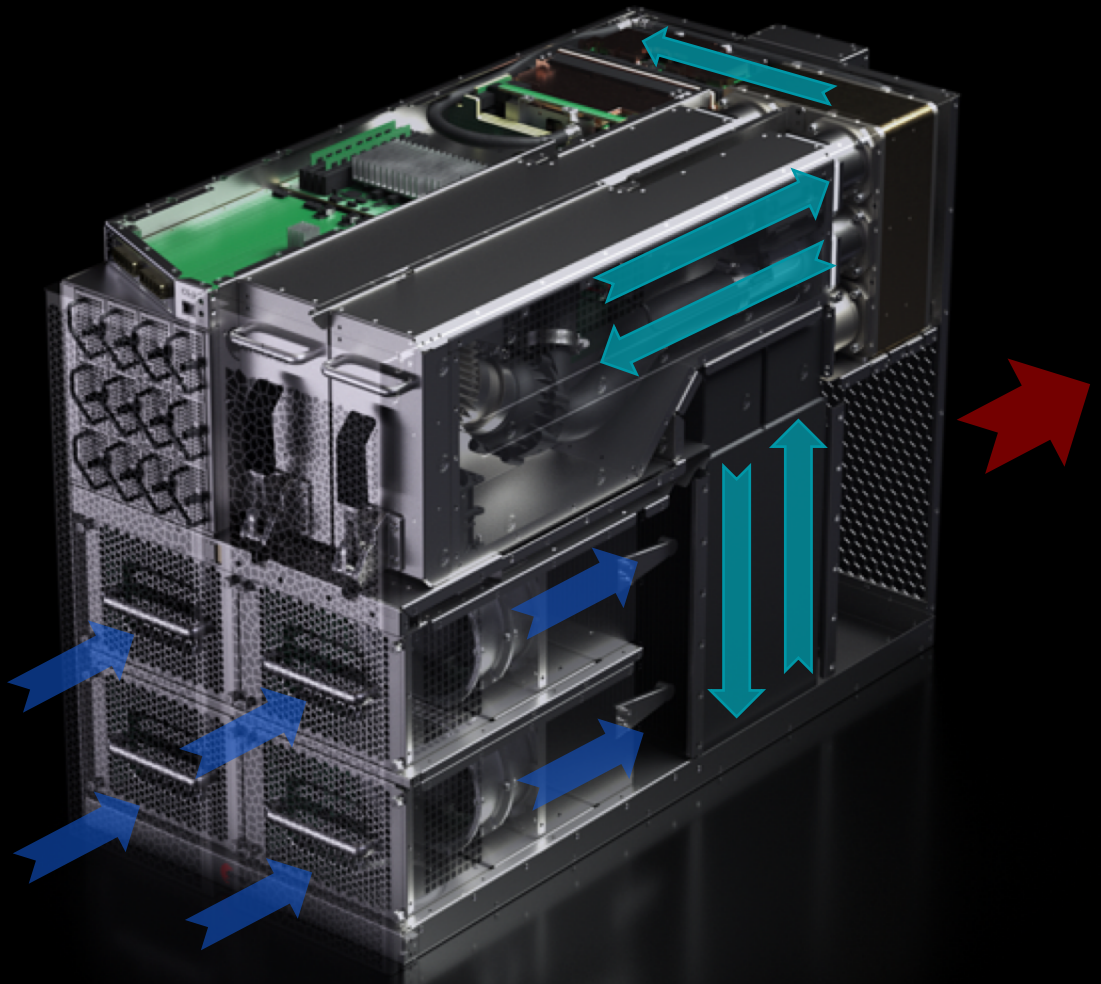
Packing the performance of a cluster into a 15RU server wasn't easy.

Required systems-level thinking, new invention and engineering for e.g.

- Packaging
- Power
- Cooling
- I/O

Let's take a peek under the hood.

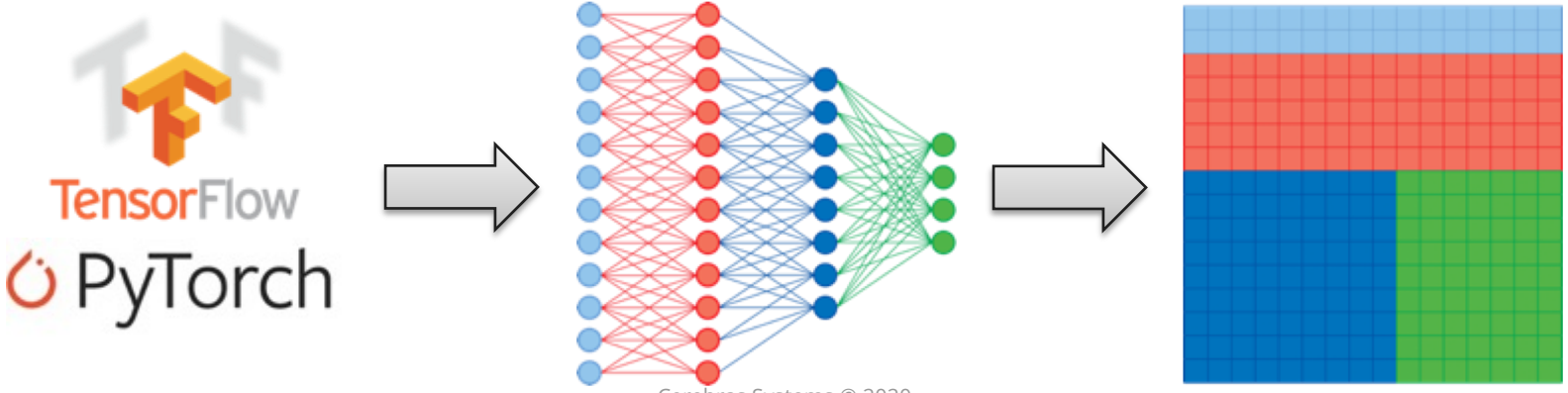
Cerebras Systems © 2020



Cerebras Systems © 2020

Programming the Wafer Scale Engine

- Neural network models expressed in common ML frameworks
- Cerebras interface to framework extracts the neural network
- Performs placement and routing to map neural network layers to fabric
- The entire wafer operates on the single neural network



Cerebras Systems © 2020

Users program transparently in ML Frameworks

Users define model_fn and input_fn **as normal**

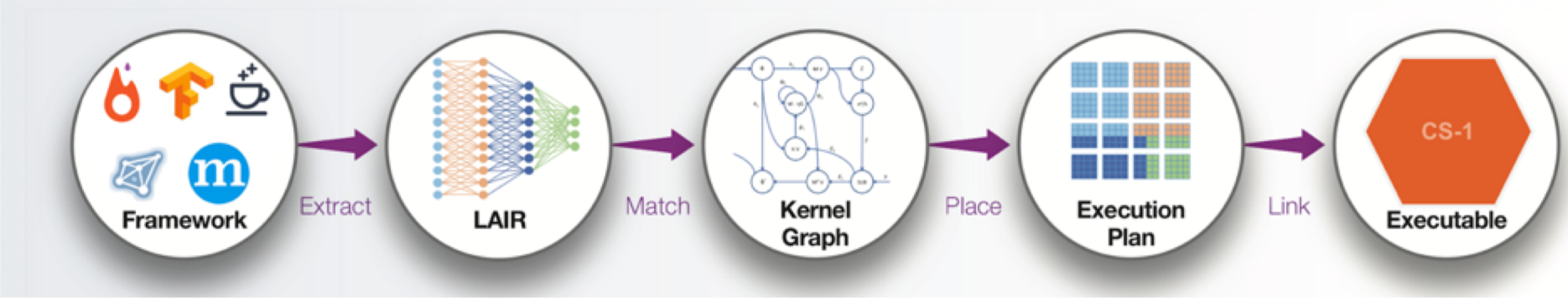
We integrate with frameworks at the **interface level**

No need to manually extract graphs or call lower-level CS-1 interaction APIs

Cerebras Systems © 2020

The Cerebras Software Platform

Our software stack makes the Wafer-Scale Engine easy to use:

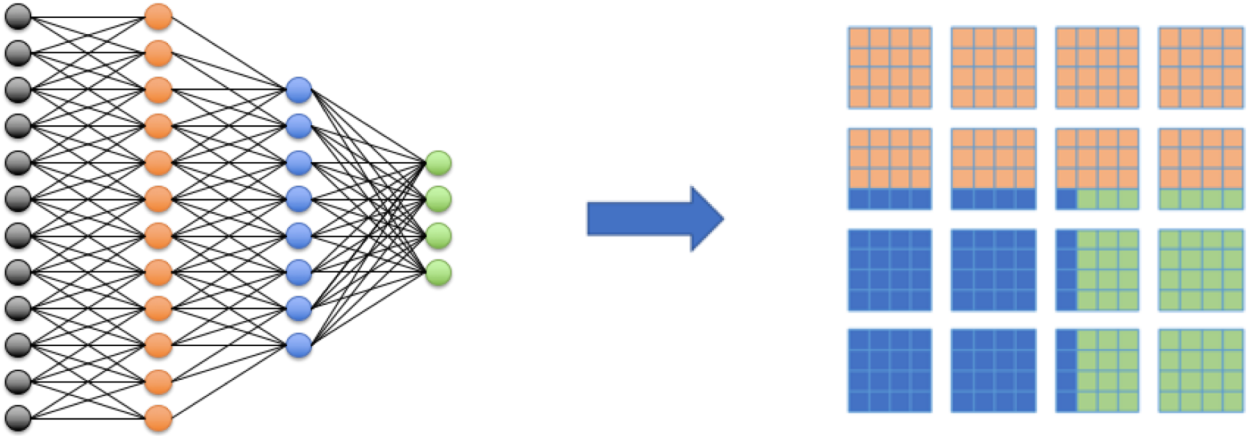


→ Programmable with today's ML frameworks
→ Library of high performance DL ops
→ Customizable and extensible for other applications with flexible lower level APIs

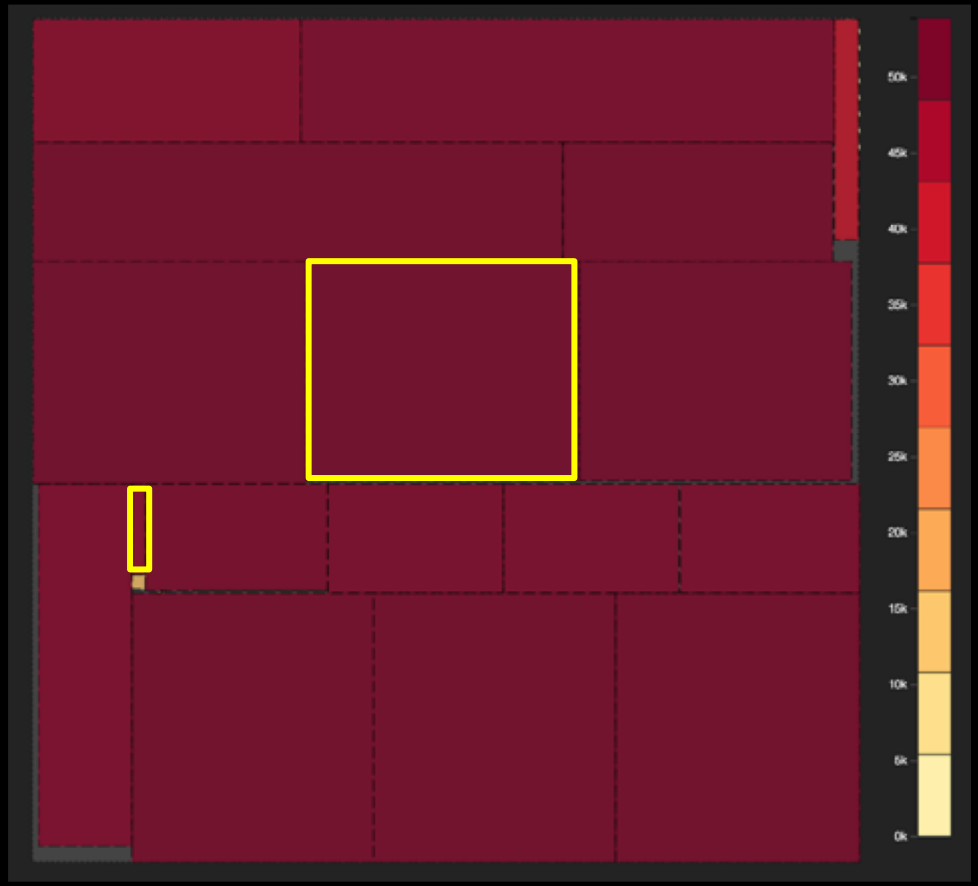
Cluster-scale AI with the programming ease of a single node

cerebras

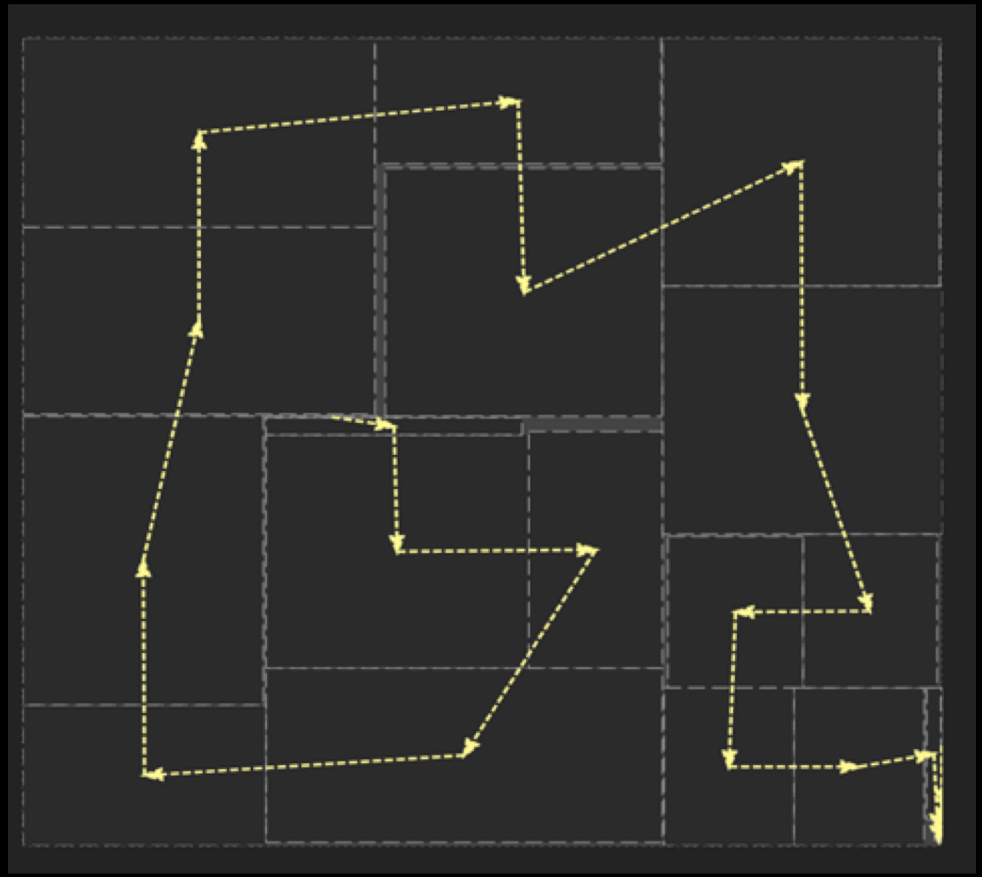
Cerebras Systems © 2020



Cerebras Systems © 2020



Cerebras Systems © 2020



Cerebras Systems © 2020

Value to users

Training time reduced from months to minutes, from weeks to seconds
→ 1,000s of new hypotheses tested in the same time period

Enable 1,000x more data in a training sets
→ More data in less time improves results

Reduce latency for inference by 1,000x
→ Latency shrinks from milliseconds to microseconds

Explore networks and methods not possible on GPUs
→ Larger deeper networks, extraordinarily sparse networks, very wide shallow networks, etc.

Already seeing orders of magnitude performance gain across multiple models + industries, e.g. web, research, pharma, finance, health & medicine.

cerebras

Cerebras Systems © 2020

WSE advantages for DL workloads

Compute is…

- **Flexible**, every core is programmable individually
- Optimized for **linear ops** (mul+add)
- **Operates on tensors**
- Dataflow architecture: ops triggered by data

➢ No wasted silicon/power
➢ Flexible support existing & future DNNs
➢ Standard DL ops accelerated out of the box
➢ Can exploit sparsity in models and data

Memory is…

- **Large, high-bandwidth, distributed, local**

➢ Utilization doesn't depend on batch size

Fabric is…

- **High bandwidth and low-latency for local communication**
- Fully **configurable**

➢ Distributed (including model parallel) training is easy

cerebras

Cerebras Systems © 2020

When do you need a CS-1?

If your **training or inference jobs are too slow and require large-scale distributed setup** today (e.g. you train your models on a cluster today).

When your **model is large, and you suffer from low utilization on GPUs due to small batch sizes** or when model no longer converges with larger effective batch size due to data parallel training.

When your **model or input data is sparse**.

Cerebras Systems © 2020

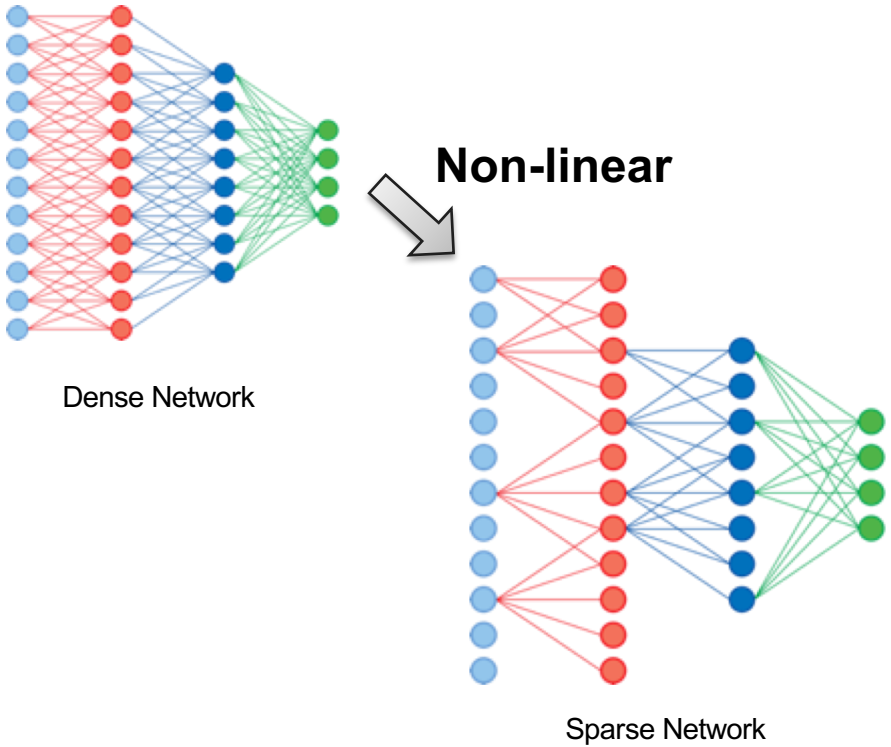
Translating Sparsity into Training Performance

Large number of zeros in neural network

- Nonlinears create activation sparsity
- Mul-Add by zero does not change the result

Kernels designed for sparsity

- Harvest natural sparsity in neural network
- Induce sparsity when not naturally occurring



Cerebras Systems © 2020

Unlock the future of AI

Brute-force scaling parameter count is the historical path to better models

Difficult to sustain growing compute and memory requirements with traditional systems

Algorithmic innovations give us more efficient models
These are promising, but challenge existing hardware

CS-1 supports both approaches to unlock future AI:
- **Extreme scale with fewer nodes**
- **Flexible compute for smarter, efficient models**

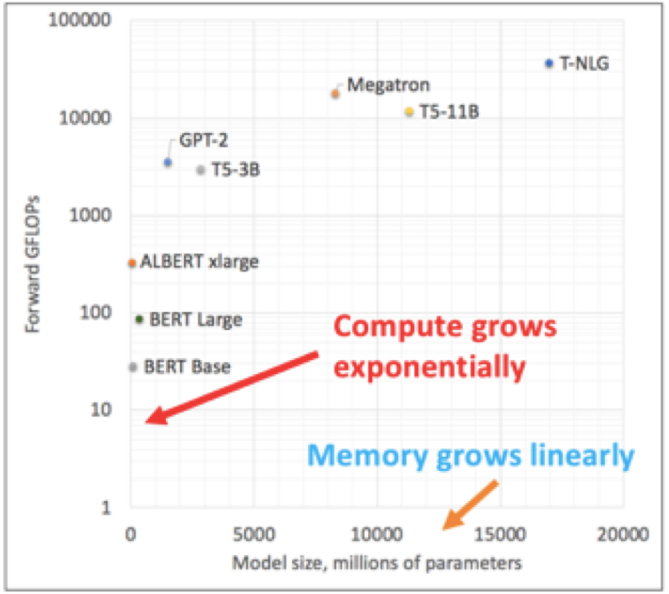
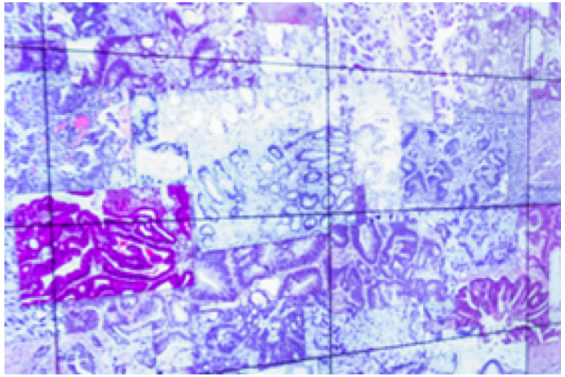


Figure. Scaling parameter count. Challenge: while memory grows linearly, compute grows exponentially.

Cerebras Systems © 2020

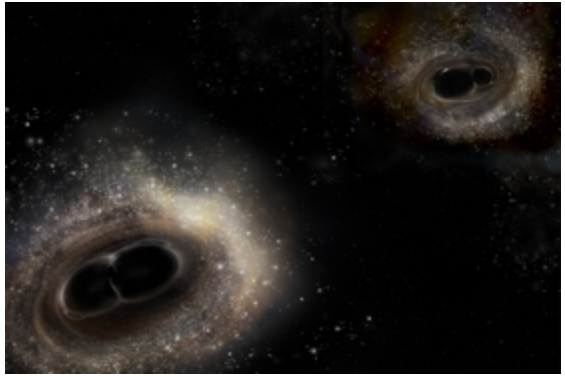
DoE national lab deployments

Cancer Treatment



ANL: Cerebras WSE guides drug treatment choices for tumors based on world's largest cancer treatment database.

Black Hole Physics



ANL: Gravitational Wave Parameter Estimation for Binary Black Hole Mergers

Lassen Supercomputer

LLNL: CS-1 + Lassen supercomputer to accelerate tightly-coupled AI + HPC.

Cerebras Systems © 2020

Thank you

Pleased to introduce Cerebras and CS-1 to ATPESC.

We built this system to **accelerate research and applied AI** for users like you, by orders of magnitude beyond legacy solutions.

We are seeing awesome demand and acceleration already;
Can't wait to see what comes next.

Thank you! Welcome questions or comments :)

Cerebras Systems © 2020